

# Network Biology Approach to Complex Diseases

## LECTURE 4. Disease Heterogeneity

Teresa Przytycka  
NIH / NLM / NCBI



# Challenges in Modeling of disease heterogeneity

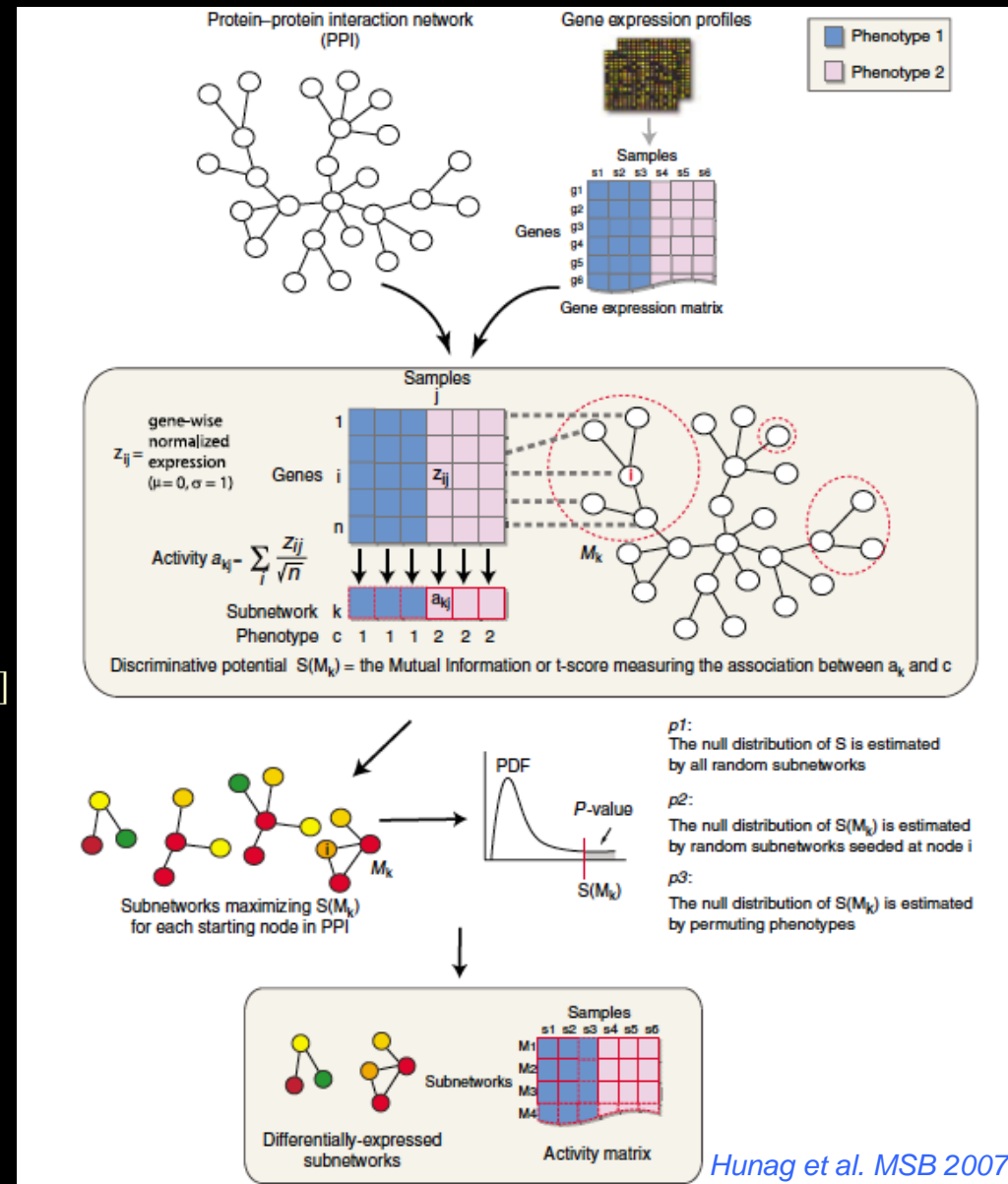
- Capturing important similarities without losing important differences
  - Examples – classification based on treatment response helps optimizing treatment options (applicable also in the absence of understanding of molecular mechanism )
- Understanding molecular underpinnings of the differences
  - Subtype specific drug design
  - Need to zoom on putative causes of differences/similarities

# Modeling Disease Heterogeneity

- **Supervised classification** discrete labels are provided by assigning a phenotype (e.g. metastatic/non-metastatic) and gene expression or other molecular measurements are taken as classifying features
  - Machine learning approaches such as random forest, SVM, etc. can be applied and will *not be discussed*
  - Network based classification (Chuang MSB 2007 and other)
- **Non supervised classification**
  - clustering using particular feature (e.g. gene expression)
  - Integrative/multi-feature clustering (example iCluster Shen et. al. PloS One 2012)
- **Network based mixture models**
  - Cho and Przytycka RECOMB 2012 / NAR 2013

# Revisiting Chuang et al. Network based classification of breast cancer metastasis

- For each gene compute activity score :
  - Normalize gene expression
  - Compute activity scores  $a_{kj}$  by averaging over genes in the subnetwork; discretize this value
- Score candidate subnetwork  $M_k$  using mutual information between value of  $a$  and phenotype
 
$$S(M_k) = \sum_{x \text{ value of } a} \sum_{y \text{ phenotype}} p(x,y) \log[ p(x,y)/p(x)p(y) ]$$
- Search for most discriminative subnetworks (greedy search)



# Comments on extensions/ modifications

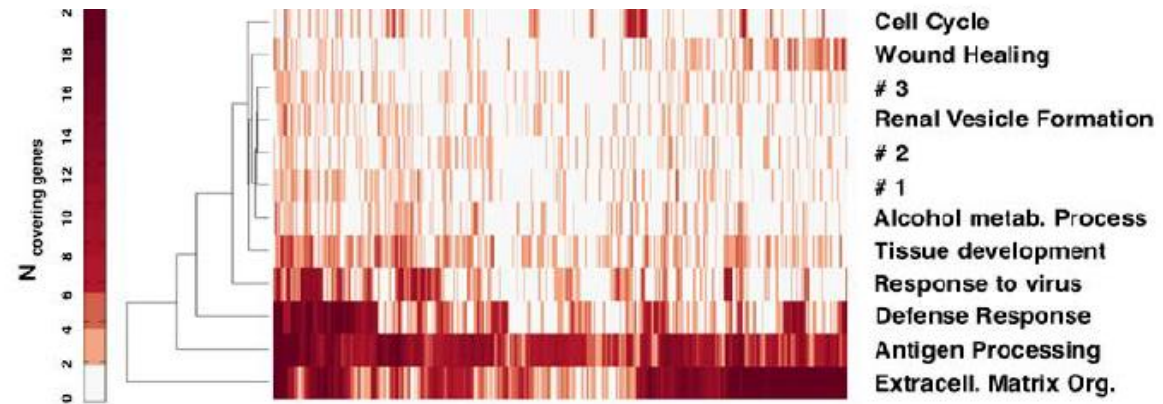
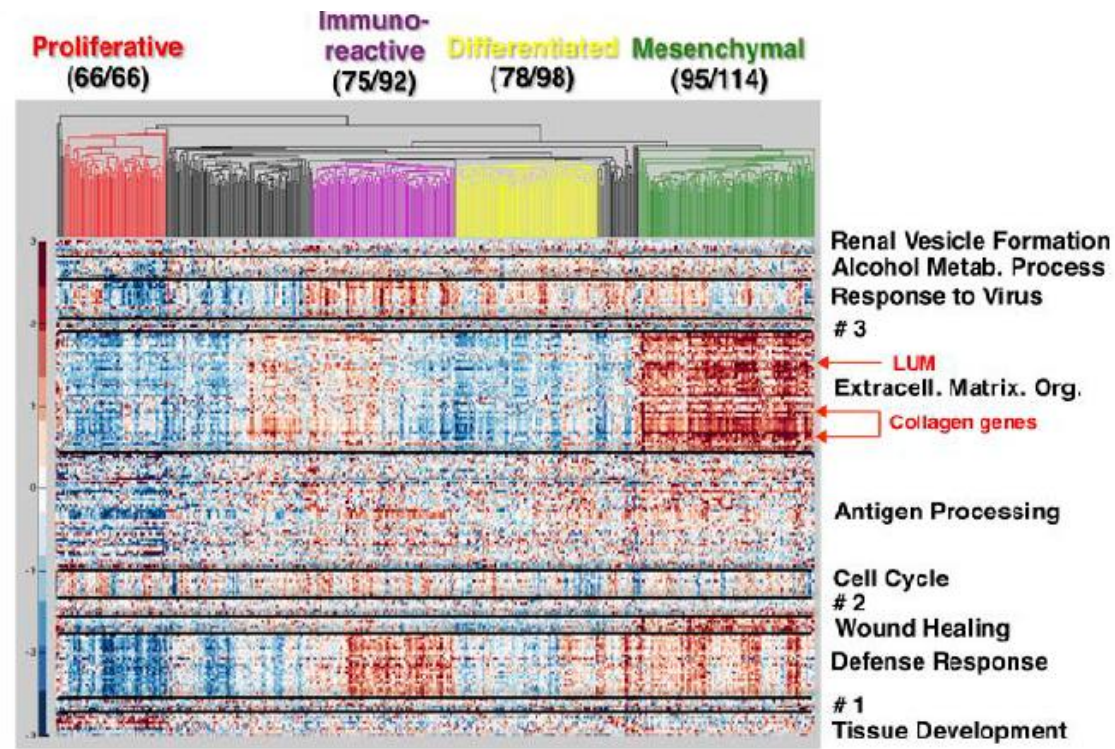
- For separating into two subclasses any method that identifies dys-regulated pathways can be used using by taking of the subtypes in place of “normal”:
  - jActive, DEGAS, module cover (already discussed)
  - Chowdhury S.A., Koyutürk M. Identification of coordinately dysregulated subnetworks in complex phenotypes. Pacific Symposium on Biocomputing 2010:133-144.
- Finding discriminative subnetworks optimally Dao et al, Bioinformatics (ISMB 2011)
  - Use color coding paradigm to find optimal subnetworks efficiently

# Non-Supervised classification

- hierarchical clustering
- positive matrix factorization
- other clustering techniques
- Integrative Clustering (iCluster)

Integrative Subtype Discovery in Glioblastoma Using iCluster  
Shen *et.al.* PloS ONE 2012

Module Cover  
TCGA Ovarian  
Cancer

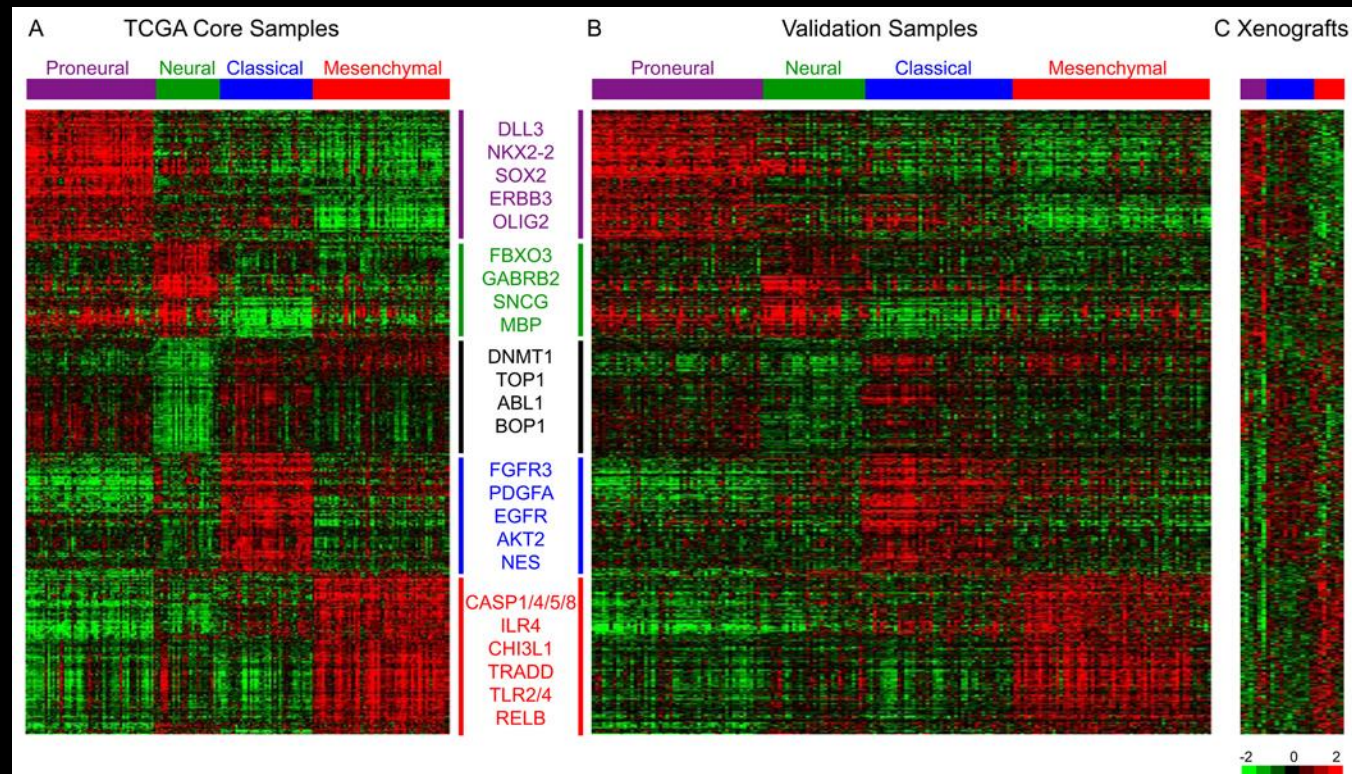




# Example – expression based clustering

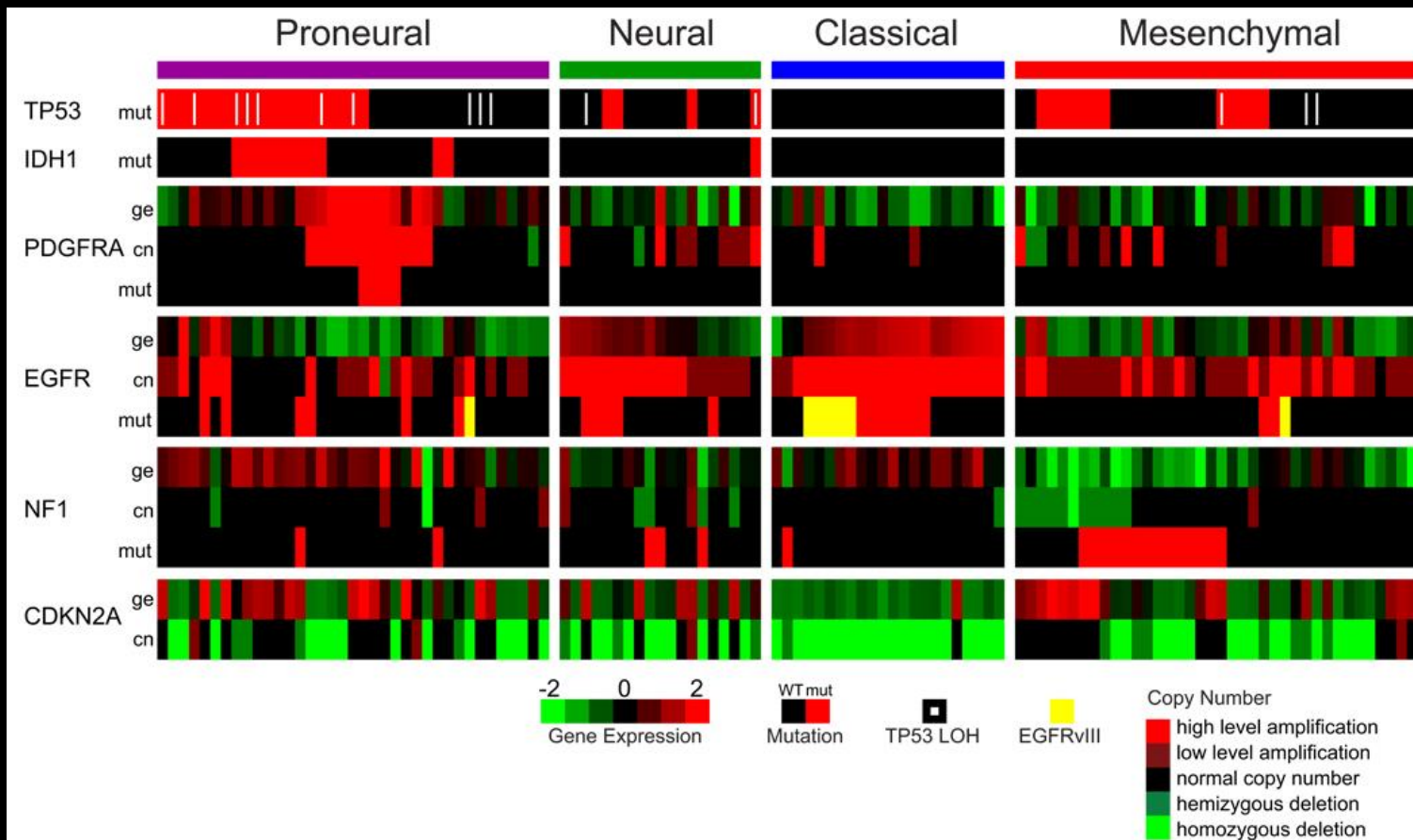
An integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR and NF1

Roel G.W. Verhaak, et al. Cancer Cell 2011





# Step 2: Identify important signatures



# Classification into subtypes is a reasonable but not perfect approach for several reasons:

- Expectation of clearly defined subgroups might not be realistic
- Difficulty in capturing underlying genotype-phenotype relation

# Key features of our approach

## Generative Topic Model

1. Our model is a meta-model that summarizes the results of 1,000 different models.
2. In each model we assume
  1.  $k$  subtypes
  2. each patient is a mixture of these subtypes
  3. each subtype is defined by distribution of features
  4. patients with similar phenotypes have similar explanatory features

# Phenotypic and explanatory features

## Phenotypic features:

Survival time

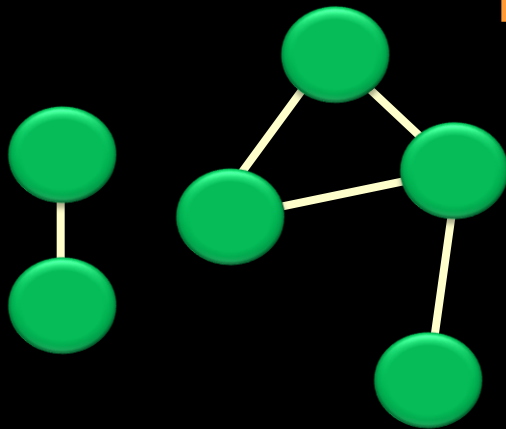
Response to drugs,.....

Gene expression profile

## Explanatory features

- mutations, CNV, micro RNA level;
- Epigenetic factors,
- Sex, environment ....

## Patient graph



*Nodes* – patients

*Edges* – phenotypic similarities

## Key idea

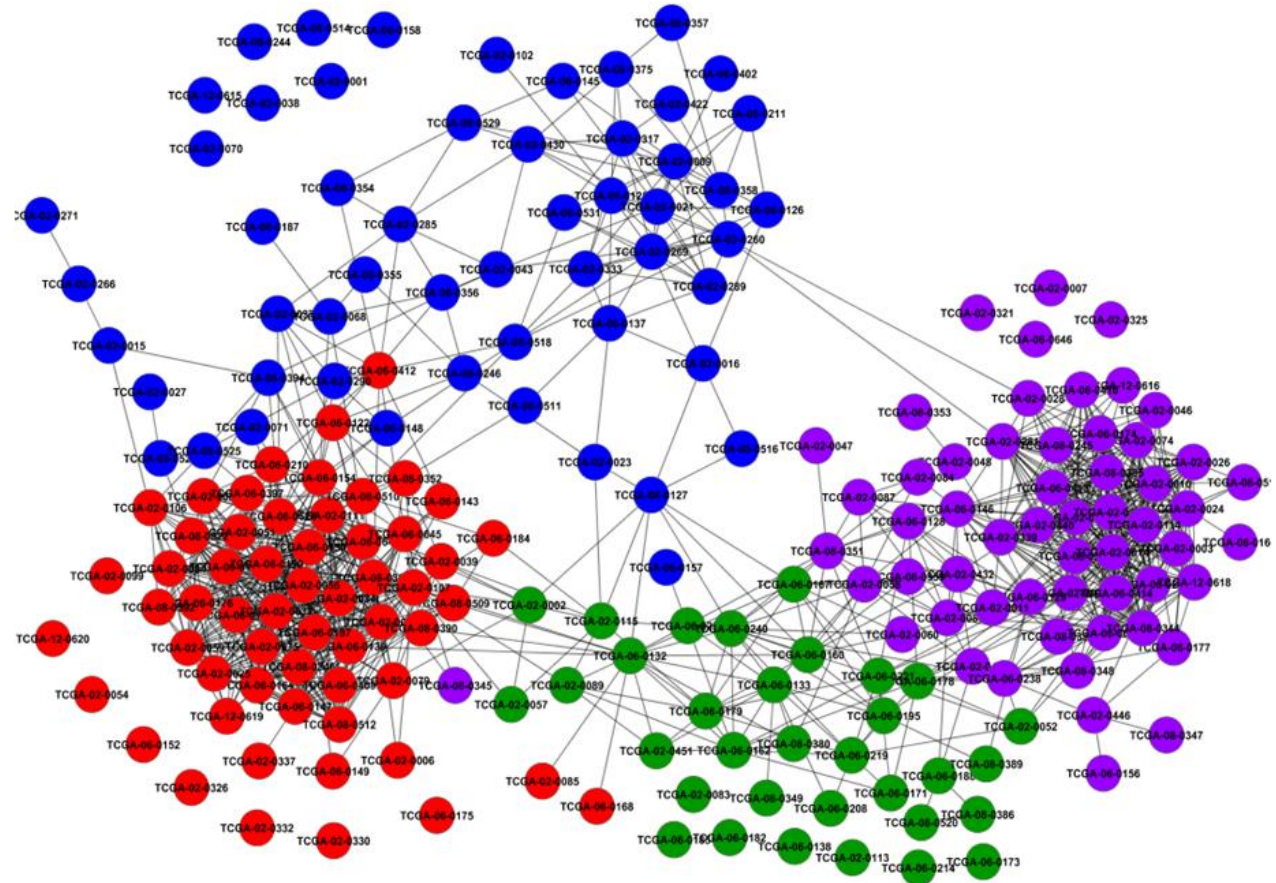
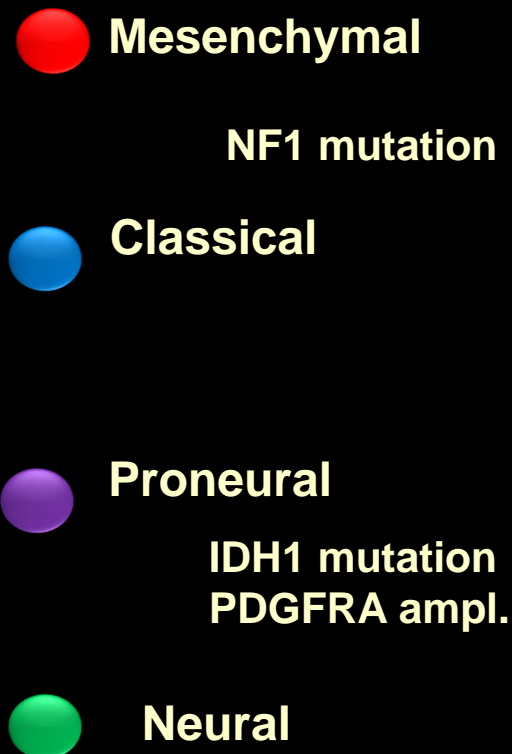
**neighbors in patient network should have similar explanatory features**

**Topic model:** Chang J, Blei DM: Hierarchical Relational Models for Document Networks. Ann Appl Stat 2010, 4(1):124-150.

# Case study of GBM (Glioblastoma Multiforme)


Varhaak et al.  
Classification

patient network for GMB



# Key features of our approach

## Generative Topic Model

1. Our model is a meta-model that summarizes the results of 1,000 different models.
2. In each model we assume 
  1. k subtypes and
  2. each patient is a mixture of subtypes
  3. Each subtype is defined by distribution of features
  4. Patients with similar phenotypes have similar explanatory features



## Step 2. 1 Assuming k subtypes, generate feature distribution for k subtypes

### Subtype I

EGFR_A	0.45
NF1_M	0.37
PTEN_M	0.21
TP53_M	0.11

...

### Subtype II

PDGFRA_A	0.51
IDH1_M	0.29
TP53_M	0.17
miR-9_H	0.11

...

### Subtype III

miR218_H	0.35
CDK2_D	0.22
SHC1_M	0.14

...

### Subtype IV

EGFR_A	0.47
CDKN2B_D	0.36
EGFR_M	0.19
miR195_H	0.05

...

- All features discretized
- one random variable per each gene and per each type of a genetic variation observed in this gene (amplification and deletion having two different variables, all mutations treated with one variable).
- For microRNA under expression as two different types of alterations where variable indicates if the expression is more than 1 or 2 standard deviations from the mean microRNA expression
- $i^{\text{th}}$  aberration in  $p^{\text{th}}$  patient corresponds to a discrete random variable  $g_{p,i}$

## Step 2. 1 Assuming k subtypes, generate feature distribution for k subtypes

### Subtype I

EGFR_A	0.45
NF1_M	0.37
PTEN_M	0.21
TP53_M	0.11
...	

### Subtype II

PDGFRA_A	0.51
IDH1_M	0.29
TP53_M	0.17
miR-9_H	0.11
...	

### Subtype III

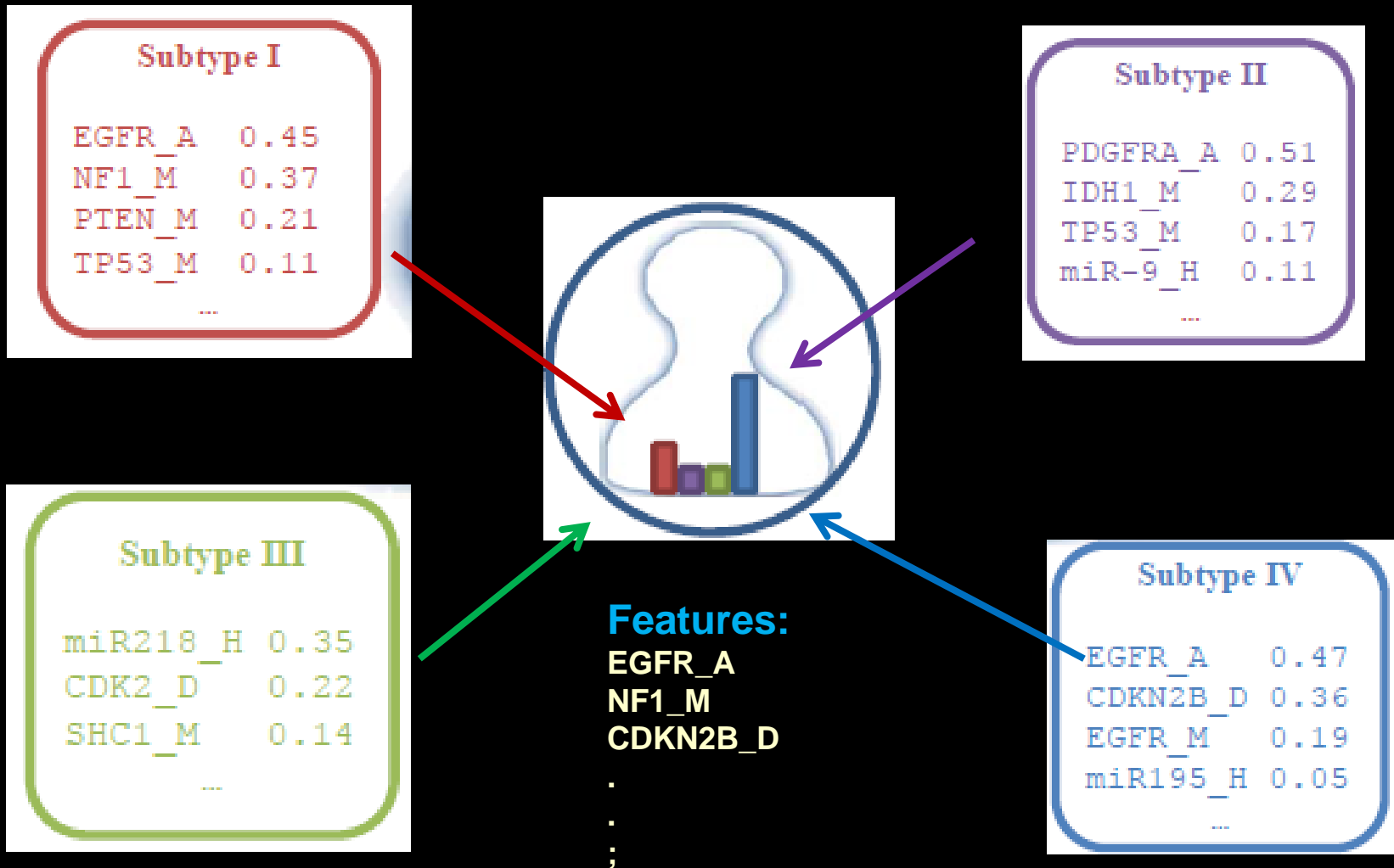
miR218_H	0.35
CDK2_D	0.22
SHC1_M	0.14
...	

### Subtype IV

EGFR_A	0.47
CDKN2B_D	0.36
EGFR_M	0.19
miR195_H	0.05
...	

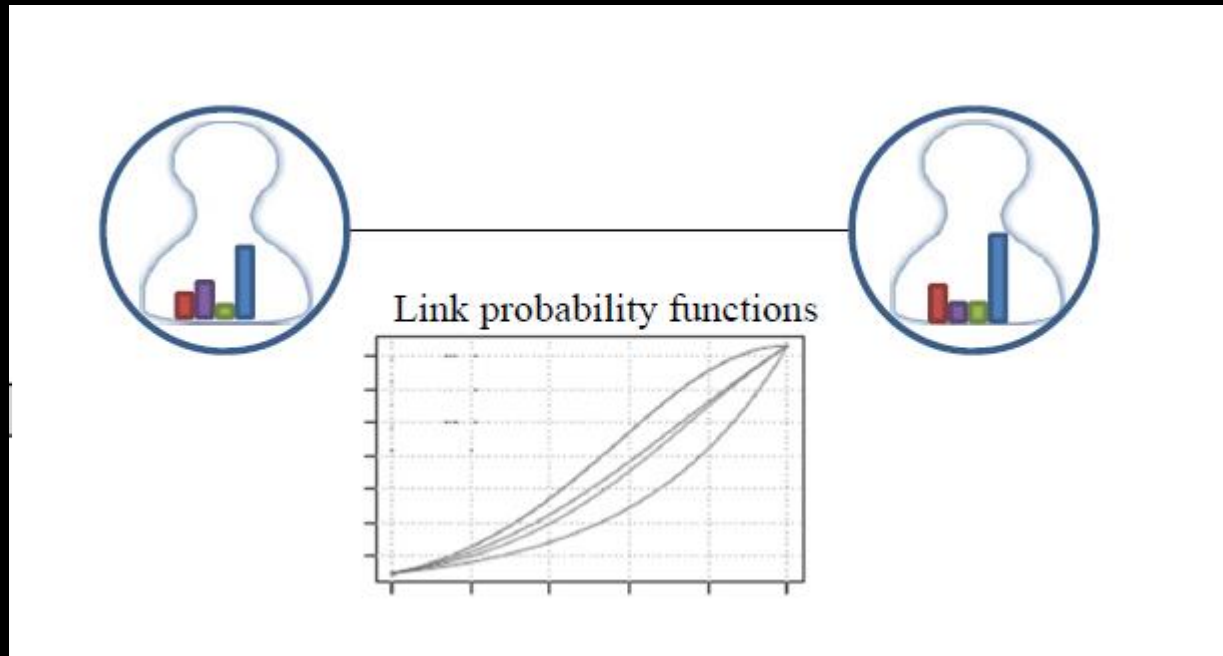
Each disease subtype  $\beta_k$  is defined as a distribution over the genomic aberrations.

## Step 2.2 .Based on patient's features represent each patient as mixture of the subtypes



First, for each patient  $p$ , draw subtype proportions  $\theta_p$  from the  $K$ -dimensional Dirichlet distribution

## Step 2.3 Generate edges based on similarity of subtype mixtures



Patient network is described by  $P^2$  binary random variables  $I_{p,p'}$ , where  $I_{p,p'}$  is set to 1 if there is a link between patients  $p$  and  $p'$ .

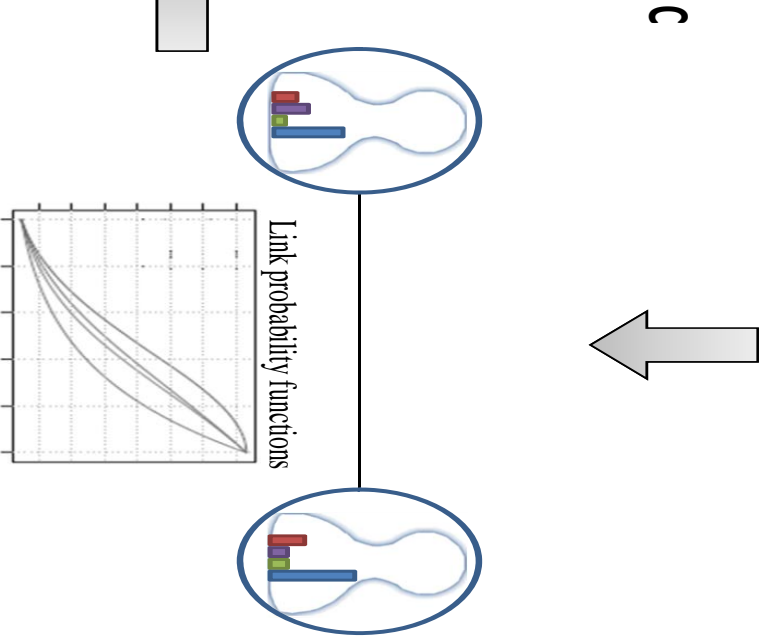
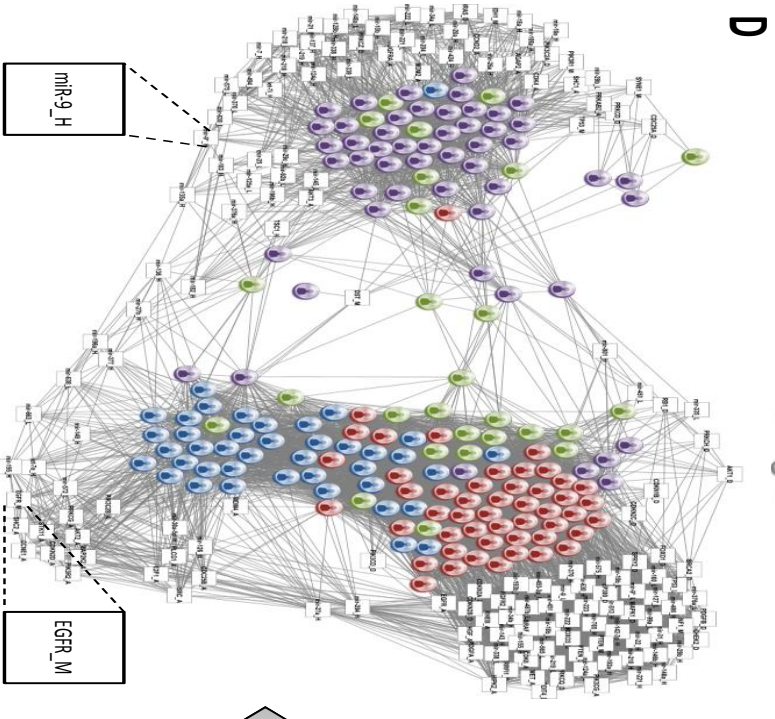
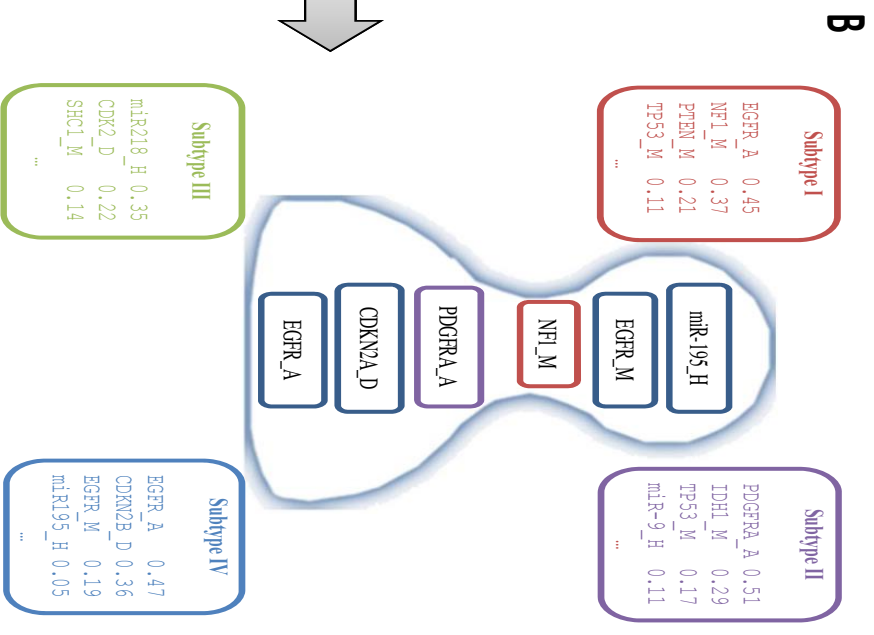
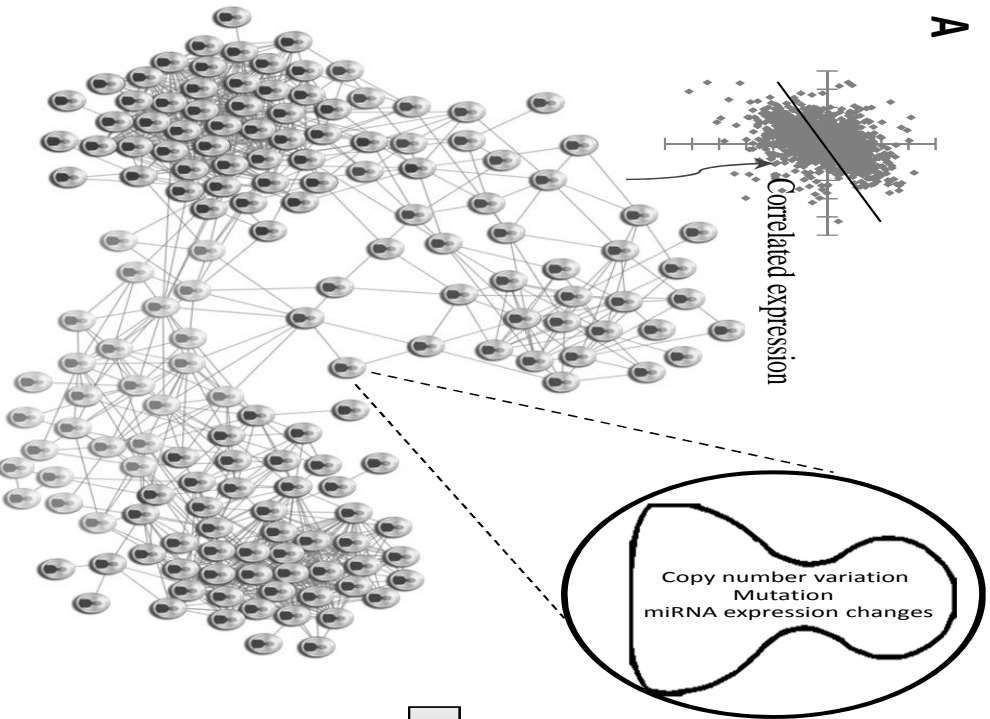
**Optimize parameters to maximize likelihood of the patient-patient network**

Chang J, Blei DM: Hierarchical Relational Models for Document Networks. Ann Appl Stat 2010, 4(1):124-150.

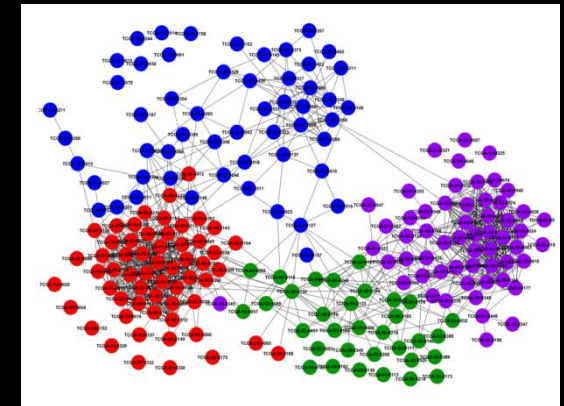
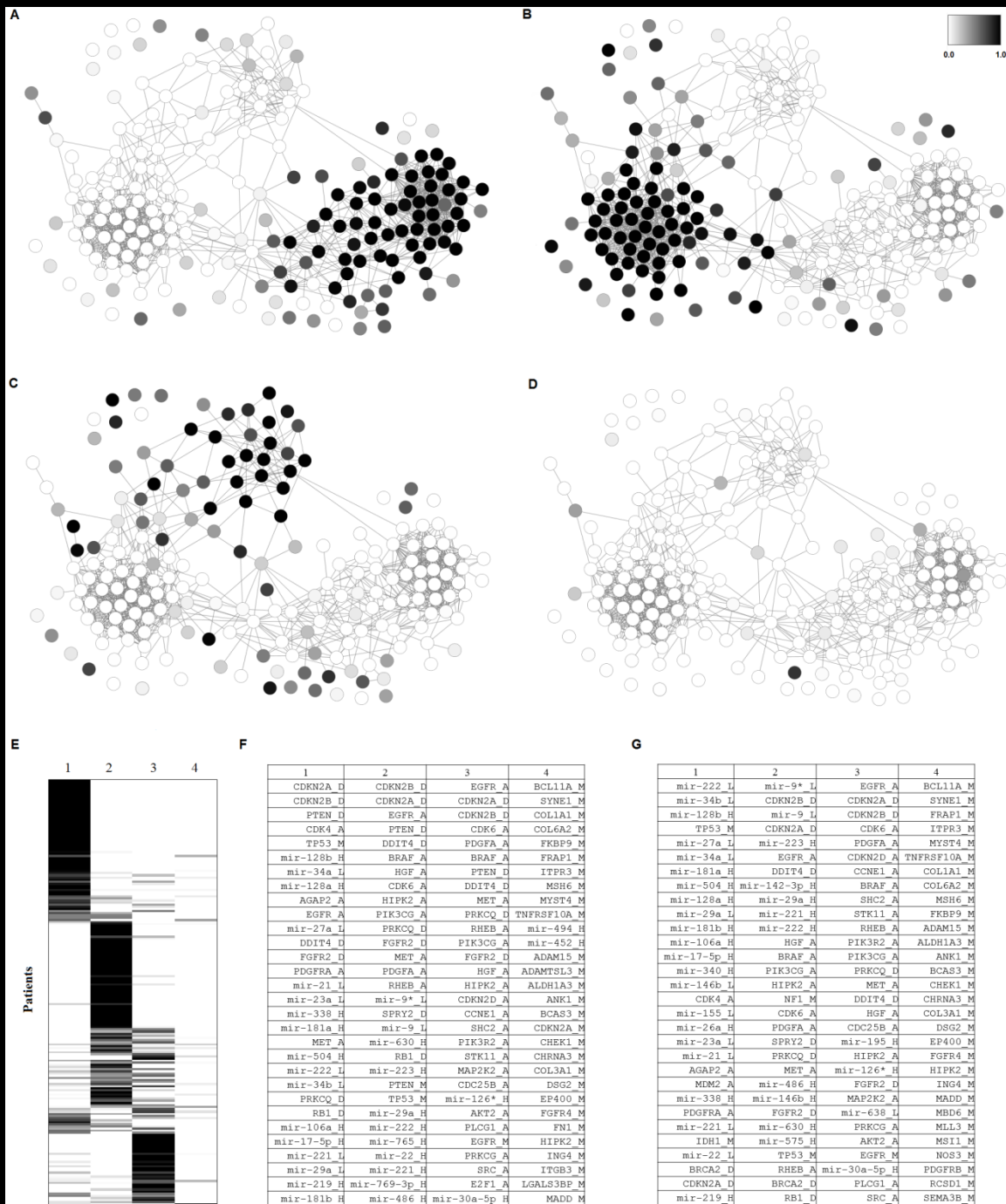
The observed patient network is assumed to be generated by the following hierarchical sampling process.

- First, for each patient  $p$ , draw subtype proportions  $\theta_p$  from the  $K$ -dimensional Dirichlet distribution.
- For each genomic factor  $g_{p,i}$ , draw the latent subtype assignment  $z_{p,i}$  from the multinomial distribution defined by  $\theta_p$  and randomly choose a genomic factor from the corresponding multinomial distribution.
- for each pair of patient  $(p, p')$  draw the binary link variable  $l_{p,p'}$  from the distribution defined by the link probability function  $\psi$ . This function is dependent on the inner product of two vectors of subtype assignments  $z_p$  and  $z_{p'}$  that generated their genomic aberrations.





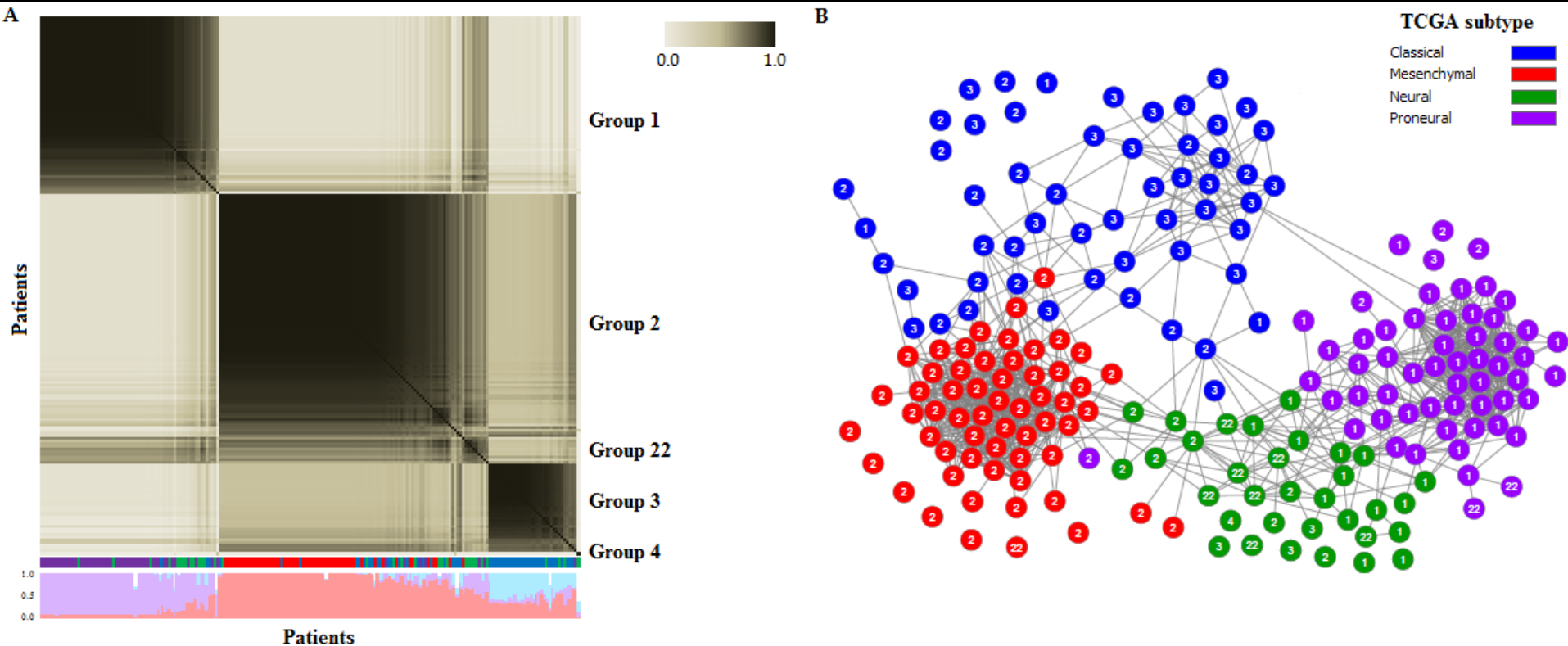
# Visualization of a sample individual model



# Summarizing the results of 1,000 models with respect to three aspects:

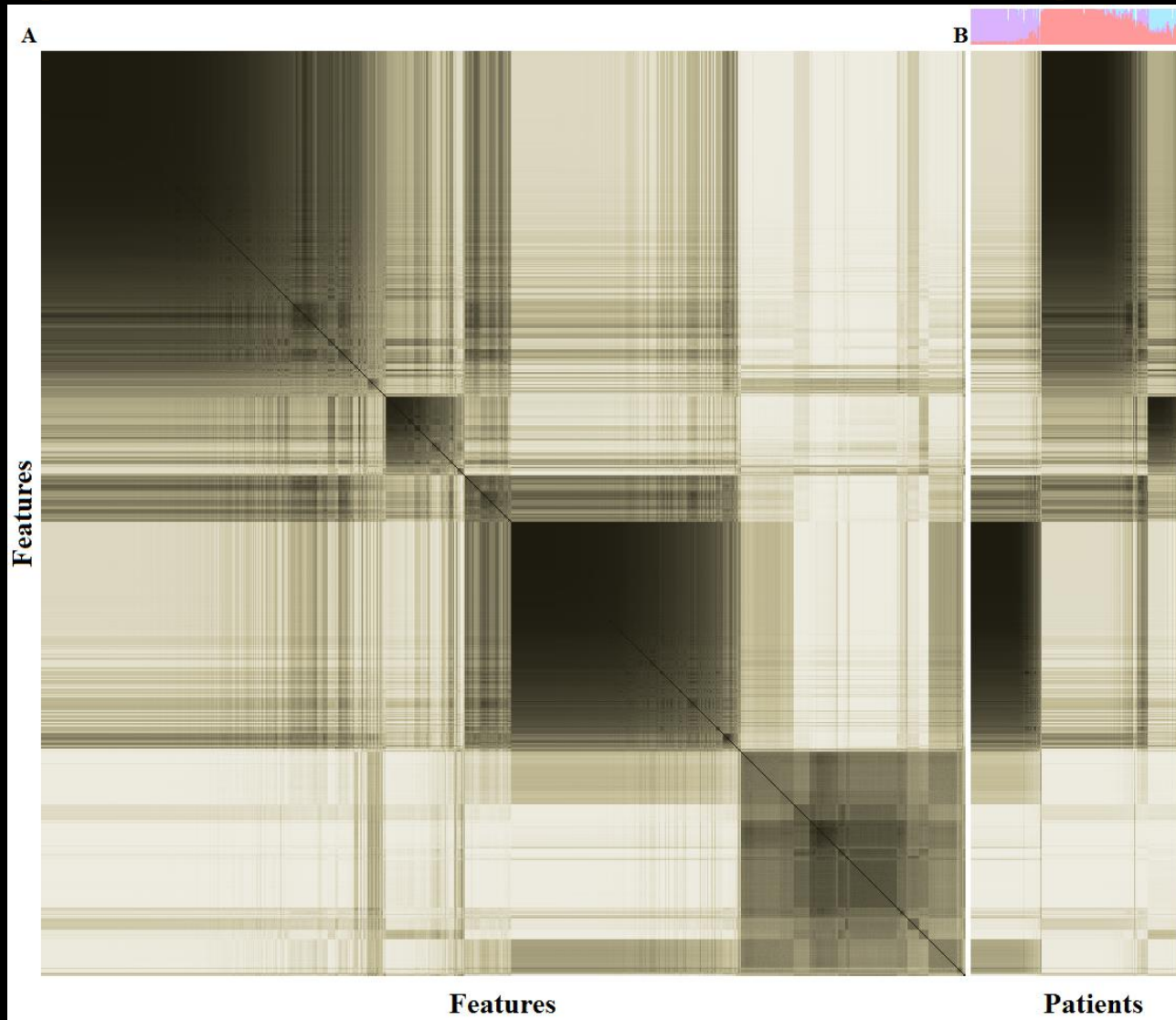
- Relation between patients
- Relation between features
- Relation between features and patients

# Patient-patient relationship



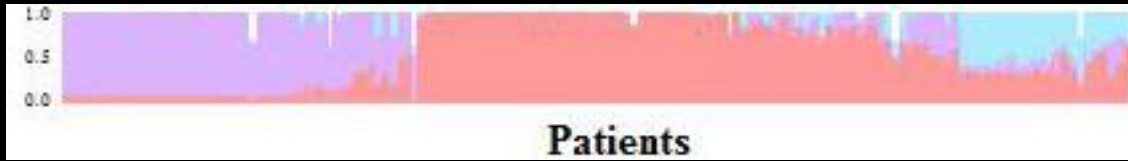
**Observation:** No separate Neural group  
(setting larger k did not change it)

# Feature-feature and patient-feature relationships

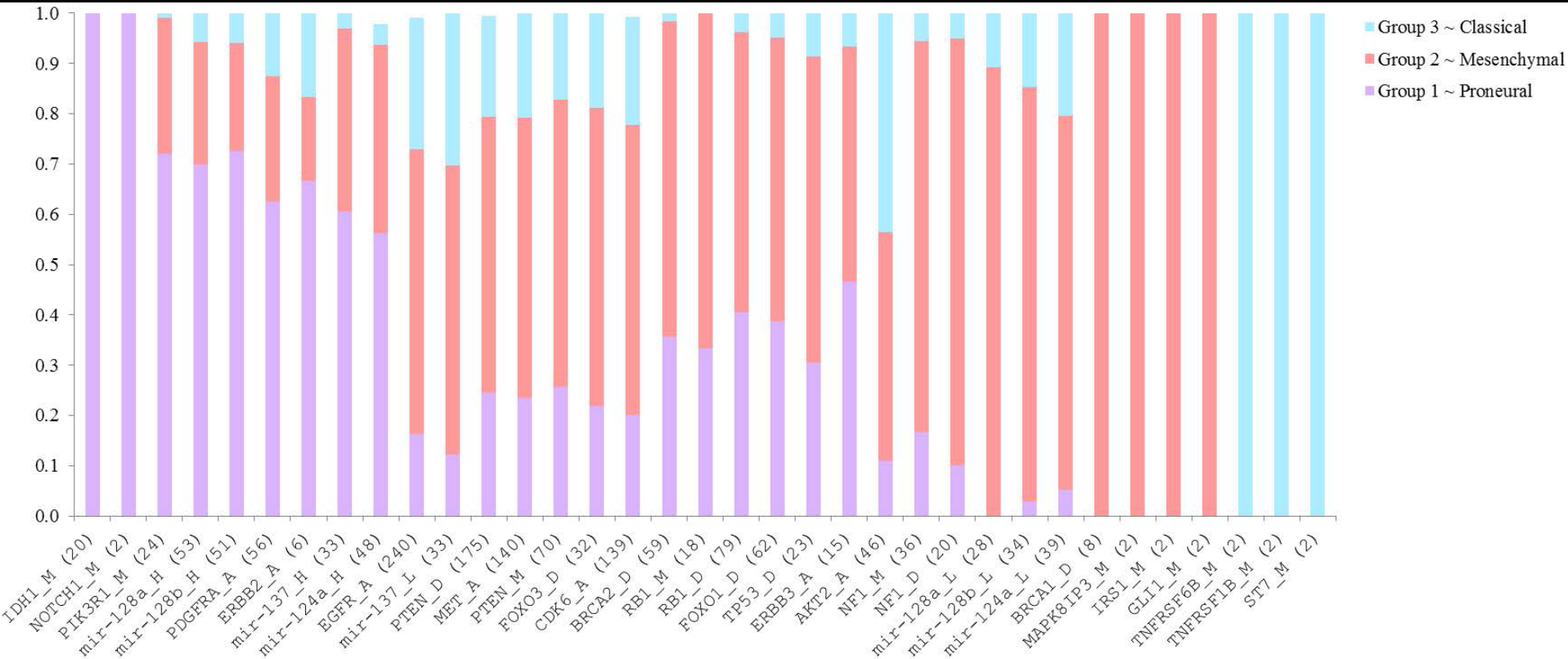




# Probabilistic subtype assignment



## Selected cancer related features





# Challenges in Modeling of disease heterogeneity

- Subtyping methods allow for capturing important similarities without losing important differences
- Mixture models – capturing overlapping subtypes
- A preferred approach should link causes to effects  
i.e. capture *genotype-phenotype relation* - *topic model*